BIOINFORMATICS APPLICATIONS NOTE

Vol. 1 no. 1 2003
Pages 1–2

MedPost: A Part of Speech Tagger for BioMedical Text

L. Smith[1], T. Rindflesch[2] and W.J. Wilbur[1]

[1] Computational Biology Branch, National Center for Biotechnology Information, 8600 Rockville Pike, Bethesda, MD, 20894, USA and [2] Cognitive Sciences Branch, Lister Hill Center for Biomedical Communication, 8600 Rockville Pike, Bethesda, MD, 20894, USA

ABSTRACT

Summary: We present a part-of-speech tagger that achieves over 97% accuracy on MEDLINE abstracts.

Availability: Software, documentation, and a corpus of 5 700 manually tagged sentences is available at ftp://ftp.ncbi.nlm.nih.gov/pub/lsmith/medpost.tag.gz

Contact: lsmith@ncbi.nlm.nih.gov

INTRODUCTION

MEDLINE is a bibliographic database of publications in health sciences, biology and related fields. It currently contains over 12 million records, and nearly 7 million include the abstract. The NCBI PubMed web site[†] provides an interface for searching MEDLINE and retrieving documents in several different formats. There is a growing amount of NLP research using biomedical text, especially MEDLINE abstracts, to improve access to the literature (information retrieval), to build databases of knowledge (information extraction), and to perform automated reasoning with biomedical knowledge (knowledge discovery). This requires increasingly effective computer comprehension of language, the starting-point for which is part-of-speech tagging, or determining the syntactic function of words in text.

The value of part-of-speech tagging degrades rapidly as the error rate increases. For example, even an error rate as low as 4% corresponds approximately to one error per sentence, which may severely limit the number of sentences that can be successfully analyzed. And taggers developed for general text do not perform this well when applied to the text of MEDLINE. This may be due to its specialized vocabulary, and perhaps also to its narrower range of stylistic variation. For example, we found nearly 57.8% of token types in MEDLINE, did not occur in either the Brown corpus (3) or the AP corpus (1988/1989 version) (this was based on 92.7% of the most common tokens in each corpus). Our tagger was developed to meet the need for a high accuracy part-of-speech tagger trained on the MEDLINE corpus.

USAGE

The *medpost* program can be run on most Unix operating systems with standard utilities (*gunzip*, *tar*, *make*, *gcc*, *perl*, *nroff*). Instructions for installing the program are contained in the file *INSTALL.medpost* which can be found in the distribution, and details on running the program can be found in a *man* page provided.

The program will currently accept text for tagging in either the native MEDLINE format or the XML format, both available as save options in PubMed. In addition, it will accept a simpler “ITAME” format that allows for text (with optional title and identifier) from any source to be tagged. The tagger segments input text into sentences and outputs each token with a part-of-speech tag separated by an underscore. For example, this is the result of tagging sentence number 9 from the MEDLINE abstract with PMID 1847596,

> Surprisingly_RR ,_, NO3-_NN inhibited_VVD the_DD rate_NN of_II K+_NN swelling_VVGN by_II 82_MC %_SYM ._.

By supplying a command line option, the tagger will automatically translate the output to either Penn treebank tag set (3) or the SPECIALIST lexicon tag set (5). Here is the same sentence above, after translation to the Penn treebank tag set,

> Surprisingly/RB ,/, NO3-/NN inhibited/VBD the/DT rate/NN of/IN K+/NN swelling/NN by/IN 82/CD %/SYM ./.

DETAILS

The MedPost tag set consists of 60 part-of-speech tags listed in table 1. It was derived from the Penn treebank tag set (3), a subset of the UCREL tag set (4), and a generalization of the SPECIALIST lexicon tag set (5). Our goal was to make the tags as unambiguous as possible, to limit their number, and to enable easy and unambiguous

[†] see http://ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed

CC	coordinating conjunction (17/991)	RR	adverb (17/651)	VHG	participle *having* (0/0)
CS	subordinating conjunction (10/120)	RRR	comparative adverb (1/1)	VHI	infinitive *have* (0/5)
CSN	comparative conjunction (*than*) (2/56)	RRT	superlative adverb (1/14)	VHN	participle *had* (0/0)
CST	complementizer (*that*) (3/122)	SYM	symbol (0/289)	VHZ	3rd pers. sing. *has* (0/28)
DB	predeterminer (0/7)	TO	infinitive marker *to* (3/159)	VVB	base form lexical verb (21/209)
DD	determiner (25/2089)	VM	modal (1/112)	VVD	past tense (64/306)
EX	existential *there* (2/19)	VBB	base *be, am, are* (1/147)	VVG	present part. (15/144)
GE	genitive marker *'s* (0/12)	VBD	past *was, were* (0/453)	VVI	infinitive lexical verb (9/127)
II	preposition (27/3470)	VBG	participle *being* (0/5)	VVN	past part. (60/815)
JJ	adjective (64/2302)	VBI	infinitive *be* (0/35)	VVZ	3rd pers. sing. (7/133)
JJR	comparative adjective (3/63)	VBN	participle *been* (0/31)	VVNJ	prenominal past part. (32/322)
JJT	superlative adjective (0/13)	VBZ	3rd pers. sing. *is* (0/162)	VVGJ	prenominal present part. (21/135)
MC	number or numeric (21/970)	VDB	base *do* (0/4)	VVGN	nominal gerund (44/152)
NN	noun (97/6344)	VDD	past *did* (0/16)	(	left parenthesis (0/456)
NNP	proper noun (18/30)	VDG	participle *doing* (0/0)	)	right parenthesis (0/463)
NNS	plural noun (42/2014)	VDI	infinite *do* (0/0)	,	comma (0/963)
PN	pronoun (3/124)	VDN	participle *done* (0/1)	.	end-of-sentence period (0/1000)
PND	determiner as pronoun (29/66)	VDZ	3rd pers. sing. *does* (0/5)	:	dashes, colons (0/115)
PNG	genitive pronoun (0/89)	VHB	base *have* (5/40)	``	left quote (5/10)
PNR	relative pronoun (7/126)	VHD	past *had* (0/45)	´´	right quote (4/13)

Table 1. The MedPost part-of-speech tag set. The number of errors per number of occurrences is given for each tag in the 1 000 sentences of the test set. Overall, the tagger achieved an accuracy of 97.43% on 26 566 tokens, with 582 sentences tagged without any errors and 261 tagged with a single tagging error.

translation to the Penn treebank and SPECIALIST lexicon tag sets.

To test and train the tagger, 5 700 sentences (155 980 tokens) were selected randomly from various subsets of MEDLINE and manually tagged. The authorities for deciding membership in word classes were (1; 2). Because of the prominence of molecular biology in bioinformatics research, most of the sentences in the corpus (used for training) were selected from themes (7) focused on this domain.

Processing begins with a perl script of regular expressions to tokenize the input following the convention of the Penn treebank (3), and to locate sentence boundaries (usually periods except for decimal points and abbreviations). The tokens of each sentence are then passed to a stochastic tagger. This employs an HMM (6) where each part-of-speech tag corresponds to a state in the Markov model, and transition probabilities are estimated from tag bigram frequencies in the training set. The output probabilities of the HMM are determined for words in a lexicon assuming equal probability for the possible tags, and for unknown words, based on word orthography (*eg* upper or lower case, numerics, *etc*), and word endings up to 4 letters long. The viterbi algorithm is used to find the most likely tag sequence in the HMM matching the tokens.

We found that to achieve high accuracy tagging required a high coverage lexicon for "open class" words (nouns, verbs, *etc*). Therefore, a lexicon 10 000 open class words was created for the most frequently occurring words in MEDLINE accounting for 92.7% of its tokens. In addition, all "closed class" words (*ie* not open class) were included in the lexicon. The entry for each word in the lexicon includes a manually entered list of all possible parts-of-speech. For a small proportion of words, and for word endings of unknown words, *a priori* probabilities for each tags is given, but for most words, the tags are assumed to occur *a priori* with equal probabilitity. Despite the lack of probability information for most words, the tagger is able to achieve high accuracy by using the contextual information in the HMM to resolve ambiguities.

REFERENCES

[1] Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J. (2000) *A Comprehensive Grammar of the English Language.* Longman, London and New York.

[2] Quirk, R. et al. (eds) (2003) *Dictionary of Contemporary English.* Longman, London and New York.

[3] Marcus, M.P., Santorini, B., and Marcinkiewicz, M.A. (1994) Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics,* **19**(2), 313-330.

[4] Garside, R., Leech, G., and McEnery, A. (1997) *Corpus Annotation.* Longman, London and New York.

[5] National Library of Medicine. (2003). *UMLS Knowledge Sources, 14th Edition.*

[6] Rabiner, L.W. (1988). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proc. of the IEEE* 77, 2, 257-286.

[7] Wilbur, W.J. (2002) *A Thematic Analysis of the AIDS Literature,* Pacific Symposium on Biocomputing **7**, 386-397.